1-1-2014

Packet loss protection for interactive speech object rendering: A multiple description approach

Xiguang Zheng
University of Wollongong, xz725@uowmail.edu.au

Christian H. Ritz
University of Wollongong, critz@uow.edu.au

Packet loss protection for interactive speech object rendering: A multiple description approach

Abstract

2014 IEEE. This paper investigates the application of novel packet loss protection schemes to compress mixtures of speech sources for interactive real-time audio services such as spatial teleconferencing. Hybrid Forward Error Correction (FEC) and Multiple Description Coding (MDC) packet loss protection techniques are applied to the outputs of a psychoacoustic-based Analysis-By-Synthesis (PABS) coder designed for such applications. The protection approaches split the coder outputs into two descriptions that are separately protected using the hybrid FEC-MDC techniques. Perceptual Evaluation of Speech Quality (PESQ) measurements compare the performance of different protection schemes for a range of typical packet loss conditions. Results indicate the proposed scheme maintains the perceptual quality of the speech sources across a wide variety of packet loss conditions.

Keywords

description, multiple, rendering, object, speech, interactive, protection, loss, approach, packet

Disciplines

Engineering | Science and Technology Studies

Publication Details

X. Zheng & C. H. Ritz, "Packet loss protection for interactive speech object rendering: A multiple description approach," in 2014 IEEE China Summit and International Conference on Signal and Information Processing, IEEE ChinaSIP 2014 - Proceedings, 2014, pp. 57-61.

PACKET LOSS PROTECTION FOR INTERACTIVE SPEECH OBJECT RENDERING: A MULTIPLE DESCRIPTION APPROACH

Xiguang Zheng[1,2], *Christian Ritz*[1]

[1]ICT Research Institute/School of Electrical Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW, Australia, 2522
[2]Dolby Laboratories (Beijing), No. 1, East 3rd Ring Middle Road, Beijing, China, 100021

xzhen@dolby.com, critz@uow.edu.au

ABSTRACT

This paper investigates the application of novel packet loss protection schemes to compress mixtures of speech sources for interactive real-time audio services such as spatial teleconferencing. Hybrid Forward Error Correction (FEC) and Multiple Description Coding (MDC) packet loss protection techniques are applied to the outputs of a psychoacoustic-based Analysis-By-Synthesis (PABS) coder designed for such applications. The protection approaches split the coder outputs into two descriptions that are separately protected using the hybrid FEC-MDC techniques. Perceptual Evaluation of Speech Quality (PESQ) measurements compare the performance of different protection schemes for a range of typical packet loss conditions. Results indicate the proposed scheme maintains the perceptual quality of the speech sources across a wide variety of packet loss conditions.

Index terms - Quality of Experience, Speech Compression, Joint Source-Channel Coding

1. INTRODUCTION

A Psychoacoustic-based Analysis-By-Synthesis (PABS) approach to compress mixtures of speech signals [1] has previously been proposed for real time audio services such spatial teleconferencing, where listeners can selectively playback speech from individual participants. The compression framework ensures that the perceptual quality of individually reconstructed speech signals is maximized at low bit rates. To ensure the Quality of Experience (QoE) is maximized for these applications, the network transmission of the resulting speech mixtures should incorporate packet loss protection that is designed to maximize the perceptual quality of decoded speech signals.

Packet loss concealment (PLC) is one of the key elements in real- time applications. On the sender side, Forward Error Correction (FEC) and Multiple Description Coding (MDC) have been widely employed to increase the robustness of the signal reconstruction against the packet loss during transmission. In FEC, once a packet is lost, it can be recovered to some extent if the redundant information regarding this packet is received. FEC can be classified into two categories, namely, the Media Independent FEC (MI-FEC) and the Media Dependent FEC (MD-FEC) [2]. The MI-FEC intends to employ algebraic coding of the bitstreams, e.g. the Reed Solomon Coding, which does not require the prior knowledge about the source signal. It is usually used for high bitrate applications. The MD-FEC is based on the idea of transmitting a degraded yet still acceptable version of the source signal as the redundancy (secondary encoding). Once the source signal is lost, the secondary will be used to recover the signal. In MDC, the source signal is divided into two or more sub-streams (descriptions) which are sent through independent transmission channels. On the receiver side, the source signal can be recovered by receiving the descriptions in any combination. The source signal can be perfectly decoded if all of the descriptions are received while a degraded source signal can be recovered if less descriptions are received. Since descriptions will be sent through independent channels and any combination of received descriptions is decodable, network congestion or packet loss for some of the transmission channels will not interrupt the decoding process but only leads to a temporary degradation of decoding quality. Comparisons between FEC and MDC techniques [3], [4] indicated that one technique will outperform another under different transmission and packet loss rates.

Proposed in this paper is a packet loss concealment scheme, based on a previous theoretical investigation of the hybrid MDC-FEC approach [5], for transmitting source recoverable speech mixtures. Such mixtures could result from joint compression of speech signals from multiple participants in a teleconference. The proposed framework employs the network status parameters (i.e. packet loss rate, available data transmission rate, etc.) to adaptively choose the optimal PLC scheme from several hybrid MDC-FEC candidates ensuring optimised speech quality for a given network status parameter set. This technique was previously implemented to transmit source recoverable audio mixtures [6] where two audio mixtures are the input signals to the hybrid MDC-FEC system, with results showing superior performance in terms of predicted audio quality compared to existing approaches that use only MDC or FEC. Here, it is extended to transmit the source recoverable speech mixtures where a mono speech mixture signal is divided into two balanced MDC descriptions.

For the intended applications, it is essential that the quality of each recovered source is maximised and not just the overall mixture signal. Hence, the source recoverable speech mixture is obtained using the previously proposed Psychoacoustic-based Analysis-By-Synthesis (PABS) [1]. The PABS framework is based on exploiting sparsity of speech in the perceptual time-frequency domain, where multiple speech signals are encoded into one mono mixture signal. Using side information indicating the active speech source for each time frequency instant of the mixture enables flexible decoding of individual speech sources. For scenarios where sparsity does not hold (e.g. simultaneously active speech sources for a given time-frequency) a psychoacoustic weighting function is used to form the mixture signal such that the overall perceptual quality for each source is maximized. The mixture signal is further compressed with a standard speech coder, in this case the AMR-WB+ coder [7] and is shown to maintain high perceptual quality at a relatively low bit rate of 36 kbps [1].

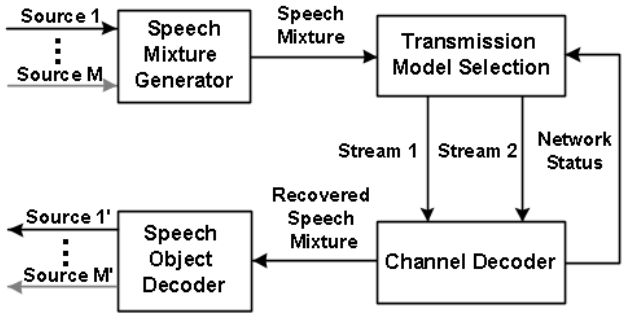


Fig.1. Proposed Joint Source-Channel Coding Framework

In this paper, the proposed hybrid MDC and FEC PLC method is combined with the PABS approach to ensure robustness to transmission through packet loss channels. An evaluation of this approach using PESQ for both anechoic and reverberant speech signals are provided to demonstrate that the proposed framework achieves significant improvement in terms of the PESQ scores during high packet loss transmission.

The remainder of the paper is organized as follows: Section 2 presents the framework and the overview of the PABS technique. Section describes the Hybrid MDC-FEC models. Objective and subjective experimental results are presented in Section 4, while conclusions are drawn in Section 5.

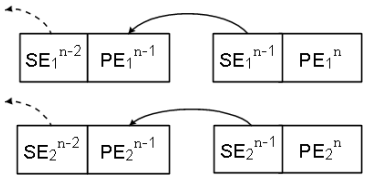
2. SPEECH MIXTURE GENERATION

The proposed framework is illustrated in Fig. 1. In the Speech Mixture Generation block, input speech objects from Source 1 to Source *M* (as shown in Fig.1) are firstly transformed into the time-frequency domain using the Short Time Fourier Transform (STFT), denoted by $S_m(n,k)$ where $1 \le m \le M$ and n and k are frame number and frequency index, respectively. Then the time-frequency speech objects will be compressed by the PABS scheme, which produces a speech mixture signal that allows for selective demixing of individual speech sources with high perceptual quality.

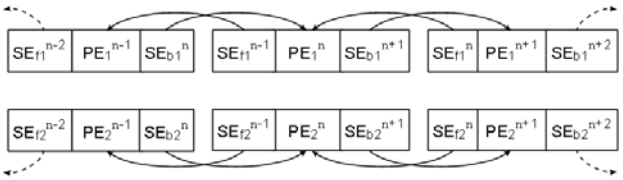
Similar to approaches used in Blind Source Separation [8], [9], the PABS approach [1] first assumes individual speech sources can be obtained from the mixture by assuming sparsity in the time-frequency domain i.e. there is only one active source dominates in each time-frequency instant. For situations where sparsity may not hold (e.g. recordings in reverberant environments), the approach uses a perceptually-weighted Analysis-By-Synthesis (ABS) optimization iteration. This approach aims to ensure that the perceptual distortion is minimised for each speech source in each frame, as measured by a psychoacoustic frequency weighing function and similar to that used successfully in speech enhancement postfilters [10]. Further details can be found in [1].

3. DYNAMIC TRANSMISSION MODEL SELECTION

In the Transmission Model Selection block, the primary and secondary audio mixtures obtained in Section 2 will be used to create different types of descriptions among MDC, FEC and previously proposed hybrid MDC and FEC models [5] according to current transmission channel parameters (i.e. available bitrates and packet loss rates).



(a) Proposed Hybrid MDC-FEC Model 1 in [5]



(b) Proposed Hybrid MDC-FEC Model 2 in [5]

Fig.2. **Hybrid MDC-FEC Models**

3.1. Hybrid MDC and FEC Models

Multiple Description Coding (MDC) [11] is a joint source-channel coding technique that divides a source bitstream into two or more independent bitstreams (descriptions) that are transmitted separately. The network congestion or packet loss for one transmission channel will not interrupt the decoding process of other channels. Full decoding quality is achieved when all descriptions are received and combined, however if only one description is received a degraded but acceptable version of the source can be obtained. For each description, an optimal transmission rate is chosen depending on the characteristics of the corresponding channel. It has been shown in [11] that balanced MDC (the transmission rate is uniform between each description) is optimal for the transmission channels having identical treatment for all packets. For the rest of this work, balanced MDC will be considered and analysed.

In Forward Error Correction (FEC) [12], lost data in the current packet (Primary Encoding) is recovered at the receiver by transmitting a degraded yet still acceptable version of the primary encoding (Secondary Encoding) in other packets. Thus, the disturbances introduced by the transmission channel can be compensated without retransmission of the primary encoding. Recently, two hybrid MDC-FEC models have been proposed in [5] and will be considered as candidate transmission models along with the standard MDC and FEC in the Transmission Model Selection block of Fig. 2. The hybrid MDC and FEC models are depicted in Fig. 2.

MDC-FEC Model 1 (Fig.2-1) implements an FEC secondary encoding with rate R_{iSE} *(i = 1 or i = 2)* to protect the primary encoding (PE) with rate R_{iPE} *(i = 1 or i = 2)* in the following packet. If the available rate for transmission channel i is R_i, $R_i = R_{iPE} + R_{iSE}$. This is indicated by the arrows in Fig.2-1. For instance, in Fig.1-1, the right package contains the secondary encoding for the n-1th primary encoding (SE_i^{n-1}), and n^{th} primary encoding (PE_i^n). The PE sections could be regarded as the MDC descriptions with coding rate R_{PE} from the two-description MDC model.

MDC-FEC Model 2 (Fig.2-2) extends MDC-FEC Model 1 by using two transmission channels to implement two SEs with rate R_{iSEf} and R_{iSEb} to protect the primary encoding with rate R_{iPE} *(i = 1 or i = 2)* in the following and previous packets, where *SEf* represents the SE of packet n that is transmitted alongside frame n-1 (a *forward* error correction scheme) and *SEb represents* the SE of packet n that is transmitted alongside packet n+1 (a *backward* error correction scheme).

Table I. Bitrate Allocation Strategies at Different Rates for Each Transmission Model

Model	Primary Encoding Rate (kbps)	Secondary Encoding Rate (kbps)	Number of Channels Employed
SDC	36	N/A	1
FEC	24	12	1
MDC	18	N/A	2
MDC-FEC-1	12	6	2
MDC-FEC-2	8	5	2

3.2. Theoretical Results for Optimised Model Selection

The theoretical rate-distortion model for MDC and FEC has been analysed in [13] based on a Gaussian source with unit variation. For the described hybrid models and the basic MDC and FEC models, it has been shown in [5] that the distortion of the transmission models can be represented as a function of the available transmission rate R and the packet loss probability p. Thus, the optimised model for a given set of R and p can be found by analysing the model with the lowest distortion criterion base on the Rate-Distortion theory [11]. It should be noted that the transmission channel parameters for practical applications could be obtained by transmission protocols such as the Real-Time Control Protocol (RTCP) [14]. Hence, for a given set of R and p feedback, the optimised model can be selected to ensure optimised quality. In this work, a similar optimised model will be derived where the input signals are real-life speech objects.

3.3. Extending the Optimised Model for Transmitting the Speech Objects Mixture

In this work, two balanced descriptions are used and each description is formed by time domain interleaving of the speech mixture created by the PABS framework. Specifically, the odd and even time domain samples of the speech mixture are extracted to form two descriptions. Thus, for a speech mixture sampled at 16 kHz, each description contains half of the samples to form a speech signal sampled at 8 kHz. This sub-sampling technique was originally proposed in [15] to form two 'self-contained' half-rate channels. Here, this idea is extended by employing the proposed hybrid MDC-FEC model. If these two descriptions are all received, perfect reconstruction can be ensured. If only one of the descriptions is received, band-limited reconstruction from the received description will be performed. A 36 kbps total transmission bitrates is employed to compress the balanced speech mixture using the AMR-WB+ Codec [7]. The bitrate allocation for each transmission model including the Single Description Coding (SDC) (i.e. transmitting the balanced descriptions via one channel without any protection scheme employed) is listed in Table I where the allocated rate for one balanced description is presented.

4. EVALUATION

4.1. Transmission Channel Model

The Gilbert Channel Model has been widely adopted to simulate the transmission characteristics. As shown is Fig. 3, it is a two state first order Markov chain process, which can be used to describe the dependencies for the packet losses.

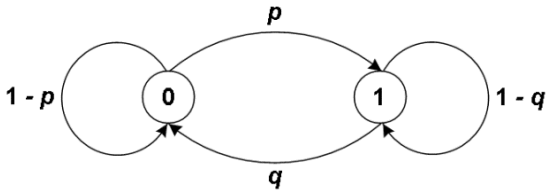


Fig.3. – Two-state Gilbert channel model. '0' indicates packet receiving while '1' represents packet loss.

Table II. Practical Gilbert Model Parameter Patterns and the Gilbert Model Parameters used in the Simulation

No.	Pattern	Counts	Selected Parameter
1	$0 < p < 0.01$, $0 < q < 0.3$	11/56	$p = 0.005$, $q = 0.1$
2	$0 < p < 0.01$, $0.3 < q < 0.7$	11/56	$p = 0.005$, $q = 0.5$
3	$0 < p < 0.01$, $0.7 < q < 1$	11/56	$p = 0.005$, $q = 0.9$
4	$0.01 < p < 0.05$, $0 < q < 0.5$	6/56	$p = 0.03$, $q = 0.25$
5	$0.01 < p < 0.05$, $0.5 < q < 1$	7/56	$p = 0.03$, $q = 0.75$
6	$0.05 < p < 1$	10/56	$p = 0.07$, $q = 0.5$

The state 0 indicates packet receiving while 1 represents packet loss. The probabilities for the transmission from state 0 to 1 and from state 1 to 0 are p and q, respectively.

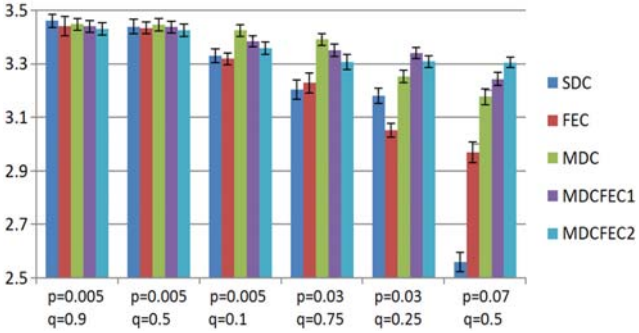
The initial probability distributions of state 0 and 1 are $S_0 = q/(p+q)$ and $S_1 = p/(p+q)$, respectively. The relationship between p and q are constrained by $1-q \geqslant p$, which can be explained since the probability of packet loss may increase but unlikely to reduce given that the previous packet is already lost. The Bernoulli Model can be considered as one of the special cases of the Gilbert Channel Model where $1-q = p$.

4.2. Testing Samples and Conditions

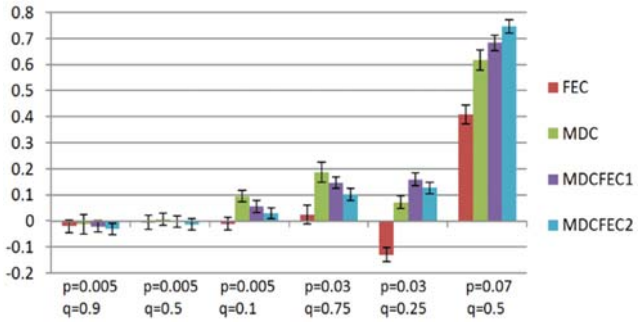
A total of 360 Sentences (sampled at 20 kHz) from the Australian National Database of Spoken Language [16] containing 36 different Australian native speakers of different ages and genders are selected as the testing database. In addition to the anechoic speech of the testing database, two reverberant conditions created from the testing database using the image method implemented through RoomSim [17] to simulate the reverberant recordings of the small (RT60 = 200 ms) and large (RT60 = 500 ms) conference room.

Speech sentences were randomly selected in groups of three and compressed using the PABS encoder to create the speech mixture signals. This process is repeated 1000 times resulting in 1000 test files for each room condition. Each speech mixture is then compressed using the proposed scheme.

The simulated transmission channel is based on the varying transmission rates and Gilbert channel parameters. After the decoding process, each recovered speech sentence is then evaluated using the PESQ [18] methodology and using the original sentence as the reference. Note that the PABS compression algorithm would also contribute to the decrease of the MOS score. Subjective listening test conducted in [1] suggests that the MOS score of the speech sources decoded from the PABS mixture signal is approximately 4.0.

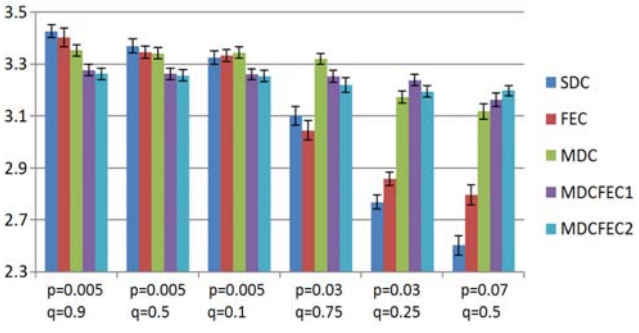


(a)

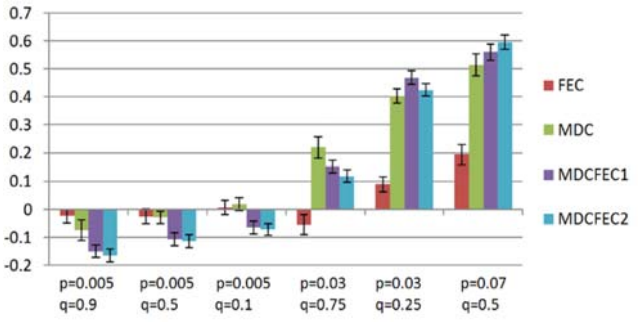


(b)

Fig. 4 PESQ (a) and PESQ Difference (b) for Hybrid MDC-FEC Anechoic Speech Mixture Transmission

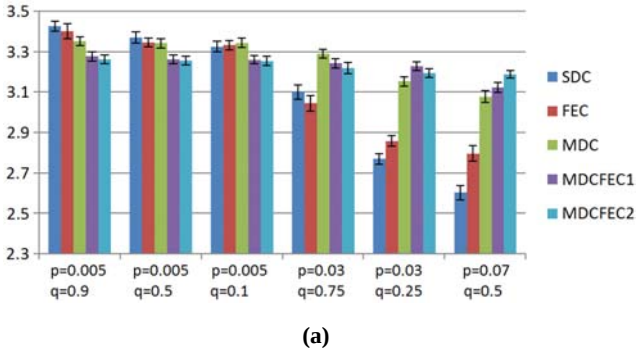


(a)

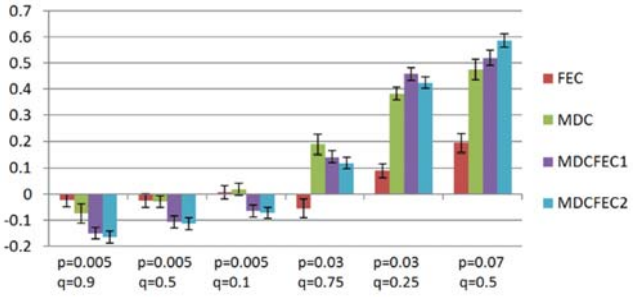


(b)

Fig. 5 PESQ (a) and PESQ Difference (b) for Hybrid MDC-FEC Reverberant (RT60 = 200 ms) Speech Mixture Transmission



(a)



(b)

Fig. 6 PESQ (a) and PESQ Difference (b) for Hybrid MDC-FEC Reverberant (RT60 = 500 ms) Speech Mixture Transmission

4.3. Evaluation Results

The PESQ evaluation results are presented in Fig. 3. The following observations can be made: (a) In low packet loss conditions (i.e. p = 0.005 and q = 0.5), no protection or simple protection strategies provides higher PESQ scores than complex protection strategies. Since for these channels packet loss is rare, the bits allocated for protection are redundant as the primary encodings are received; (b) For intermediate packet loss rates (i.e. p = 0.03; q = 0.75 or p = 0.005; q = 0.01), MDC achieves the highest PESQ scores; (c) For high level packet loss rates (i.e. p = 0.03; q = 0.25 or p = 0.09), the hybrid MDC-FEC models provide the highest performance in terms of PESQ scores (up to 0.75 higher compared to the SDC condition). The optimized models for each Gilbert channel parameter pair is given by Table II.

5. CONCLUSION

This paper has proposed the application of hybrid MDC-FEC models for packet loss protection of compressed speech mixtures obtained from the PABS framework [1]. The average PESQ for individually recovered speech signals for anechoic as well as two different reverberant environments were measured. Results show that the proposed approach significantly improves the perceptual quality of speech sources transmitted through high packet loss channels when compared to standard SDC or FEC approaches. The approach can adaptively choose the most optimal packet loss protection scheme based on predicted subjective speech quality.

ACKNOWLEDGEMENT

This work has been supported by the Australian Research Council (ARC) through the grant DP1094053.

REFERENCES

[1] X. Zheng, C. Ritz, and J. Xi, "Encoding Navigable Speech Sources: A Psychoacoustic-Based Analysis-by-Synthesis Approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 29 –38, Jan. 2013.

[2] G. Dan, V. Fodor, and G. Karlsson, "A Rate-Distortion Based Comparison of Media-Dependent FEC and MDC for Real-Time Audio," in *IEEE International Conference on Communications, 2006. ICC '06*, 2006, vol. 3, pp. 1002–1007.

[3] Z. Li, S. Bruhn, S. Zhao, and J. Kuang, "Analytical and Experimental Comparison of Packet Loss Recovery Methods Based on AMR-WB for VoIP," in *IEEE International Conference on Communications, 2009. ICC '09*, 2009, pp. 1 –6.

[4] M. R. Stoufs, A. Munteanu, J. Barbarien, J. Cornelis, and P. Schelkens, "Error protection of scalable soures: A comparative analysis of Forward Error Correction and Multiple Description Coding," in *2009 16th International Conference on Digital Signal Processing*, 2009, pp. 1–6.

[5] X. Zheng and C. Ritz, "Hybrid FEC and MDC Models for Low-Delay Packet-Loss Recovery," in *5th International Conference on Signal Processing and Communication Systems*, Honolulu, USA, 2011.

[6] X. Zheng and C. Ritz, "Packet loss protection for interactive audio object rendering: A multiple description approach," in *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2012, pp. 68 –73.

[7] T. 26. 29. 3GPP Specification series, "Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions," 2009.

[8] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings*, 2000, vol. 5, pp. 2985–2988 vol.5.

[9] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830– 1847, Jul. 2004.

[10] S. Gustafsson, P. Jax, A. Kamphausen, and P. Vary, "A postfilter for echo and noise reduction avoiding the problem of musical tones," in *, 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. Proceedings*, 1999, vol. 2, pp. 873–876 vol.2.

[11] V. K. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

[12] E. Altman, C. Barakat, and V. M. Ramos, "Queueing analysis of simple FEC schemes for IP telephony," in *IEEE INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings*, 2001, vol. 2, pp. 796–804 vol.2.

[13] Moo Young Kim and W. B. Kleijn, "Comparative rate-distortion performance of multiple description coding for real-time audiovisual communication over the Internet," *IEEE Transactions on Communications*, vol. 54, no. 4, pp. 625– 636, Apr. 2006.

[14] T. Friedman, R. Caceres, and A. Clark, "RTCP extended reports," *[Online]: http://tools.ietf.org/rfc/ rfc3611.txt*, 2003.

[15] N. Jayant, "Subsampling of a DPCM speech channel to provide two 'self-contained' half-rate channels," *Bell System Technical Journal*, 1981.

[16] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, "The Australian National Database of Spoken Language," in *, 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94*, 1994, vol. i, pp. I/97–I100 vol.1.

[17] D. Campbell, K. Palomäki, and G. Brown, "A MATLAB simulation of 'shoebox' room acoustics for use in research and teaching.," *Computing and Information Systems Journal, ISSN 1352-9404*, vol. 9, no. 3, 2005.

[18] ITU, "P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." 2001.